4Д

УДК 004.931

И.В. Дрозд, Е.В. Волченко
Государственный университет информатики и искусственного интеллекта,
г. Донецк, Украина

Метод сокращения обучающих выборок GridDC

Предложен новый метод сокращения обучающих выборок GridDC (Grid-density-center method), основанный на покрытии признакового пространства сеткой и нахождении единственного объекта клетки как объекта новой обучающей выборки. Предложен принцип формирования сетки и способы построения объектов сокращенной обучающей выборки. Для определения эффективности предложенного метода проведен сравнительный экспериментальный анализ с известными методами сокращения обучающих выборок, показавший эффективность метода GridDC.

Введение

Большинство известных алгоритмов построения решающих правил в обучающихся системах распознавания используют в качестве обучающей выборки специально отобранное подмножество обучающих объектов [1]. Примерами таких алгоритмов могут быть алгоритм Relief [2], в котором выбираются «средние» объекты выборки, и метод опорных векторов SVM [3], в котором решающее правило строится по объектам, лежащим вблизи межклассовой границы. Использование сокращенных обучающих выборок позволяет повысить скорость и качество классификации, существенно уменьшить объем хранимых данных.

Наиболее известными методами сокращения обучающих выборок являются: NNR (nearest neighbor rule), LVQ (learning vector quantization), ADM (Astrahan’s density-based method), STOLP [4], [5]. Идеей метода NNR [4] является получение минимального подмножества точек, таких, что находятся ближе всего к k-ближайшим соседним объектам. В алгоритме LVQ [4] область признаков делится на число отдельных регионов и для каждого региона строится вектор признаков нового объекта. Метод ADM [4] состоит из двух основных этапов: выбор максимальной плотности точек, основанной на локальной оценке плотности, и отсечение других точек, которые лежат рядом с выбранными точками. Идея метода STOLP [5] заключается в нахождении «напряженных» пограничных точек, на основе которых выполняется пробное распознавание всех точек обучающей выборки по правилу ближайшего соседа. Среди не верно классифицированных точек выбирается та, у которой максимальный вес, и она добавляется к набору пограничных. Рассмотренные выше алгоритмы сокращения обучающих выборок имеют общую идею, идентичную задаче кластеризации, которая состоит в разбиении исходной выборки на подмножества и их обработку. На сегодняшний день одним из наиболее эффективных современных алгоритмов кластеризации является сеточный алгоритм [6]. Основной особенностью алгоритма является переход от кластеризации отдельных объектов к обработке объектов, принадлежащих некоторой клетке сеточной структуры. Особенно эффективно данный алгоритм применяется для кластеризации выборок большого объема и позволяет выделить кластеры сложной формы. В данной работе рассматривается возможность применения сеточной структуры для задачи сокращения обучающих выборок.

Целью работы является разработка метода сокращения обучающих выборок в системах распознавания.

Постановка задачи. Пусть дано некоторое множество объектов $X=\{X_1, X_2, \ldots, X_n\}$, n – размер обучающей выборки, представленное в виде объединения непересекающихся множеств, называемых классами $X=\bigcup_{q=1}^{l} W_q$. Каждый объект описывается

системой признаков $X_i = \{x_{i1}, x_{i2}, \dots, x_{ik}\}$. Имеется конечное множество объектов, в каждом из которых известно, к какому классу он принадлежит. Необходимо построить новую сокращенную обучающую выборку объектов $X' = \{X'_1, X'_2, \dots, X'_{n'}\}$, n' – размер новой обучающей выборки.

Описание метода

В статье предлагается новый метод GridDC (Grid-density-center method), целью которого является сокращение обучающей выборки для уменьшения машинного времени на обучение. Идеей предлагаемого метода является наложение сетки на признаковое пространство, определение объектов выборки, принадлежащих каждой из клеток сетки, и их замена на объекты новой сокращенной обучающей выборки. Далее предложено пошаговое описание предлагаемого метода.

Шаг 1. Формирование сетки. Рассчитывается шаг клетки s по формуле:

$$s = \left\lfloor 1 + \frac{\left(\sum_{i=1}^{k} (\max\{x_i\} - \min\{x_i\})\right)^k * (\lfloor \ln(n) \rfloor - 1)}{k * \prod_{i=1}^{k} (\max\{x_i\} - \min\{x_i\})} \right\rfloor, \tag{1}$$

где $\lfloor ... \rfloor$ – оператор округления до ближайшего целого значения, $\max\{x_i\}$ – максимальное значение i-признака среди всех объектов выборки, $\min\{x_i\}$ – минимальное значение i-признака среди всех объектов выборки.

Рассчитывается плотность клетки (количество объектов, попавших в клетку) d по формуле:

$$d = \left\lfloor \frac{\prod_{i=1}^{k} (\max\{x_i\} - \min\{x_i\}) * (\lfloor \ln(n) \rfloor + 1)}{\left(\sum_{i=1}^{k} (\max\{x_i\} - \min\{x_i\})\right)^k} \right\rfloor. \tag{2}$$

Шаг 2. Формирование множества объектов клеток $G_i = \{G_{i1}, G_{i2}, ..., G_{ij}\}$, где G_{ij} – j объект i клетки.

Если в i клетке количество объектов больше или равно плотности d и все объекты клетки принадлежат к одному классу, то рассчитывается центр текущей клетки, который является объектом X'_q новой сокращенной обучающей выборки X':

$$\begin{cases} |\mathrm{Gi}| \ge d, \text{ то } X'_q = \left\{ \dfrac{b_{1,n+1} - b_{k,n}}{2}, \dfrac{b_{2,n+1} - b_{2,n}}{2}, ..., \dfrac{b_{k,n+1} - b_{k,n}}{2} \right\}, \\ |\mathrm{Gi}| < d, \text{ то для клетки объект новой выборки не строится} \end{cases} \tag{3}$$

где $b_{k,n}$ – левая граница текущей клетки, $b_{k,n+1}$ – правая граница текущей клетки.

В результате выполнения вышеуказанного алгоритма будет получена новая сокращенная обучающая выборка X'.

Анализируя предложенный метод, можно выделить следующие особенности:

1) обрабатываются только те клетки, в которых объекты принадлежат к одному классу W_j :

$$\{G_{i1}, G_{i2}, ..., G_{ip}\} \in W_j; \tag{4}$$

2) если в некоторой клетке находятся объекты, принадлежащие разным классам, то объект новой обучающей выборки не строится, поскольку такие клетки находятся на межклассовой границе и могут представлять собой шум;

3) объект новой обучающей выборки относится к тому же классу, что и объекты в клетке, по которым он был сформирован.

Способы формирования объектов сокращенной обучающей выборки

В зависимости от расположения объектов в клетке можно предложить несколько способов формирования объектов новой сокращенной обучающей выборки.

1 Значения признаков объекта новой выборки рассчитываются как координаты центра текущей клетки:

$$X'_q = \left\{ \frac{b_{1,n+1} - b_{k,n}}{2}, \frac{b_{2,n+1} - b_{2,n}}{2}, ..., \frac{b_{k,n+1} - b_{k,n}}{2} \right\}. \quad (5)$$

2 Значения признаков объекта новой выборки рассчитываются как координаты центра масс объектов текущей клетки:

$$X'_q = \left\{ \frac{\sum_{i=1}^{|Gi|} x_{1,i}}{|Gi|}, \frac{\sum_{i=1}^{|Gi|} x_{2,i}}{|Gi|}, ..., \frac{\sum_{i=1}^{|Gi|} x_{|Gi|,i}}{|Gi|} \right\}. \quad (6)$$

3 Значения признаков объекта новой выборки рассчитываются как координаты центра прямоугольника, описанного вокруг объектов текущей клетки:

$$X'_q = \left\{ \frac{\max\{x_{1,i}\} - \min\{x_{1,i}\}}{2}, \frac{\max\{x_{2,i}\} - \min\{x_{2,i}\}}{2}, ..., \frac{\max\{x_{|Gi|,i}\} - \min\{x_{|Gi|,i}\}}{2} \right\}. \quad (7)$$

Отметим, что выбор способа формирования объектов сокращенной обучающей выборки зависит от решаемой задачи. В данной статье по результатам экспериментальных исследований на тестовых данных производится выбор способа, рекомендованного к использованию для большинства прикладных задач.

Теоретические оценки метода GridDC

Основу предлагаемого метода GridDC составляет выбор крайних объектов образов в признаковом пространстве и поочередное отнесение каждого из объектов выборки к одной из клеток сетки. Исходя из этого временная сложность предлагаемого метода равна O(n).

Исходя из принципов построения сетки можно показать, что максимальное количество клеток $maxG_i$ будет равно:

$$\max \mathrm{G_i} = \left\lfloor \prod_{i=1}^{k} \left(\frac{\max\{x_i\} - \min\{x_i\}}{s} \right) \right\rfloor. \quad (8)$$

Если предположить, что объекты сокращенной выборки будут построены по всем клеткам сетки, то количество объектов новой сокращенной выборки будет равно:

$$\frac{\mathrm{n}}{\mathrm{d}}. \quad (9)$$

Таким образом, количество объектов новой выборки в общем случае существенно меньше количества объектов исходной выборки.

Экспериментальные исследования

Для оценки эффективности предложенного алгоритма GridDC был проведен сравнительный анализ с методом STOLP и ADM на серии испытаний. В качестве входных данных использовалась обучающая выборка объектов двух классов с двумя признаками, распределенными по нормальному закону. Оценка эффективности выполняется

по количеству неверных классификаций. Для построения решающего правила использовался метод потенциальных функций [7]. На рис. 1 приведена зависимость количества неверно классифицированных объектов *m* от общего количества объектов исходной выборки *n* при площади пересечения 25%.

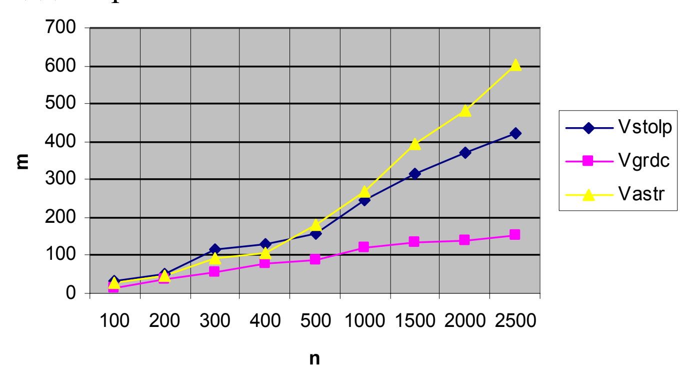


Рисунок 1 – Зависимость количества неверно классифицированных объектов от общего количества объектов исходной выборки при площади пересечения 25%

На рис. 2 приведена зависимость количества объектов построенных сокращенных обучающих выборок n' от общего количества объектов исходной выборки *n*. Данные являются усредненными по результатам 100 проведенных экспериментов. Для оценки качества классификации объектов методом GridDC с использованием предлагаемых способов формирования новых объектов сокращенной выборки была проведена сравнительная характеристика на ряде экспериментов. На рис. 3 приведена зависимость количества неверно классифицированных объектов *m* от общего количества объектов исходной выборки *n* при площади пересечения 25%, где Vgrdc1 – метод GridDC, когда значения признаков объекта новой выборки рассчитываются как координаты центра текущей клетки; Vgrdc2 – метод GridDC, когда значения признаков объекта новой выборки рассчитываются как координаты центра прямоугольника, описанного вокруг объектов текущей клетки; Vgrdc3 – метод GridDC, когда значения признаков объекта новой выборки рассчитываются как координаты центра масс объектов текущей клетки.

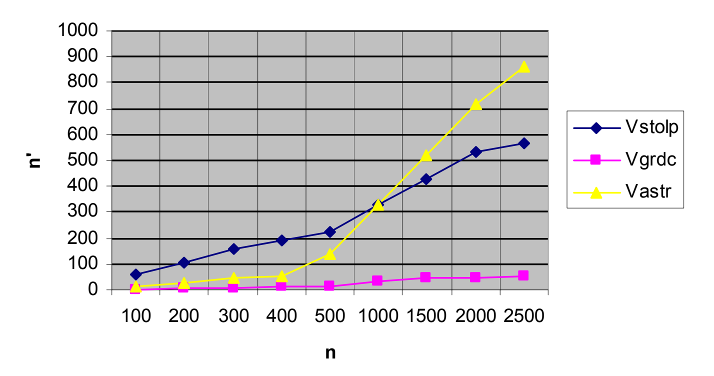


Рисунок 2 – Зависимость количества объектов построенных сокращенных обучающих выборок от общего количества объектов исходной выборки

Анализ полученных результатов позволяет сделать вывод, что при увеличении размера обучающей выборки число неверно классифицированных объектов методами STOLP и ADM значительно больше в сравнении с предложенным методом GridDC. Также показано, что размер сокращенных обучающих выборок, полученных методами STOLP и ADM, значительно больше размера выборок, полученных предложенным методом GridDC.

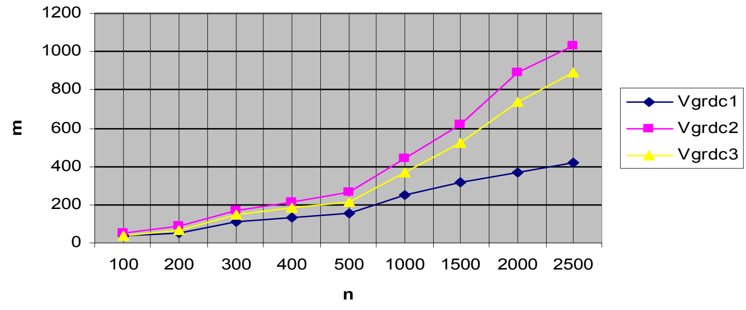


Рисунок 3 – Зависимость количества неверно классифицированных объектов от способов формирования объектов обучающих выборок

Выводы

В данной статье предложен новый метод GridDC сокращения объектов обучающих выборок. Он основывается на использовании сетчатых методов, которые до этого времени применялись только для решения задач кластеризации. Описана общая схема метода, предложен принцип расчета шага сетки и способы формирования объектов новой сокращенной обучающей выборки. Показано, что предложенный метод имеет линейную временную сложность и позволяет существенно уменьшить количество объектов в выборке. По результатам проведенных исследований показана эффективность метода GridDC в сравнении с известными методами STOLP и ADM по качеству классификации и размеру получаемых выборок.

Литература

1. Воронцов К.В. Профили компактности и выделение опорных объектов в метрических алгоритмах классификации / К.В. Воронцов, А.О. Колосков // Искусственный интеллект. – 2006. – № 2. – С. 30-33.
2. Liu H. A Selective Sampling Approach to Active Feature Selection Artificial Intelligence / H. Liu, H. Motoda, L. Yu. – 2004. – Is. 1, V. 159.
3. Vapnik V. Statistical Learning Theory / Vapnik V. – N.Y. : John Wiley & Sons, Inc., 1998. – 732 p.
4. Pal, Sankar K. Pattern recognition algorithms for data mining: scalability, knowledge discovery, and soft granular computing / Sankar K. Pal and Pabitra Mitra. – Florida : CRC Press LLC, 2004. – P. 244.
5. Загоруйко Н.Г. Методы распознавания и их применение / Загоруйко Н.Г. – М. : Сов. радио, 1972. – 206 с.
6. Куликова Е.А. Непараметрический алгоритм кластеризации для обработки больших массивов данных / Е.А. Куликова, И.А. Пестунов, Ю.Н. Синявский // ММРО. – 2009. – № 14. – С. 149-152.
7. Айзерман М. Метод потенциальных функций в теории обучения машин / Айзерман М., Браверман Э., Розоноэр А. – М. : Наука, 1970. – 384 с.

І.В. Дрозд, О.В. Волченко

Метод скорочення навчальних вибірок GridDC

Метою роботи є розробка методу скорочення навчальних вибірок у системах розпізнавання. Запропоновано новий метод скорочення навчальних вибірок GridDC (Grid-density-center method), який базується на покритті ознакового простору сіткою і знаходженні єдиного об’єкта клітки як об’єкта нової навчальної вибірки. Запропоновано принцип формування сітки і способи побудови об’єктів скороченої навчальної вибірки. Для визначення ефективності запропонованого методу проведений порівняльний експериментальний аналіз з відомими методами скорочення навчальних вибірок, що показав ефективність методу GridDC.

I.V. Drozd, E.V. Volchenko

A Method of the Reduction of the Teaching Selections GridDC

A grid-density-center method of the reduction of the teaching selections in the recognition systems is proposed. It is based on coverage of character space and on finding the unique object of a cage as object of new teaching selections. Principle of forming of the grid and methods of the construction of the objects of brief teaching selection are offered. For calculation of the efficiency of the offered method a comparative experimental analysis is conducted with the known methods. The analyses have shown that the method increases accuracy the classification and decrease the length of the teaching selections.

Статья поступила в редакцию 05.07.2010.